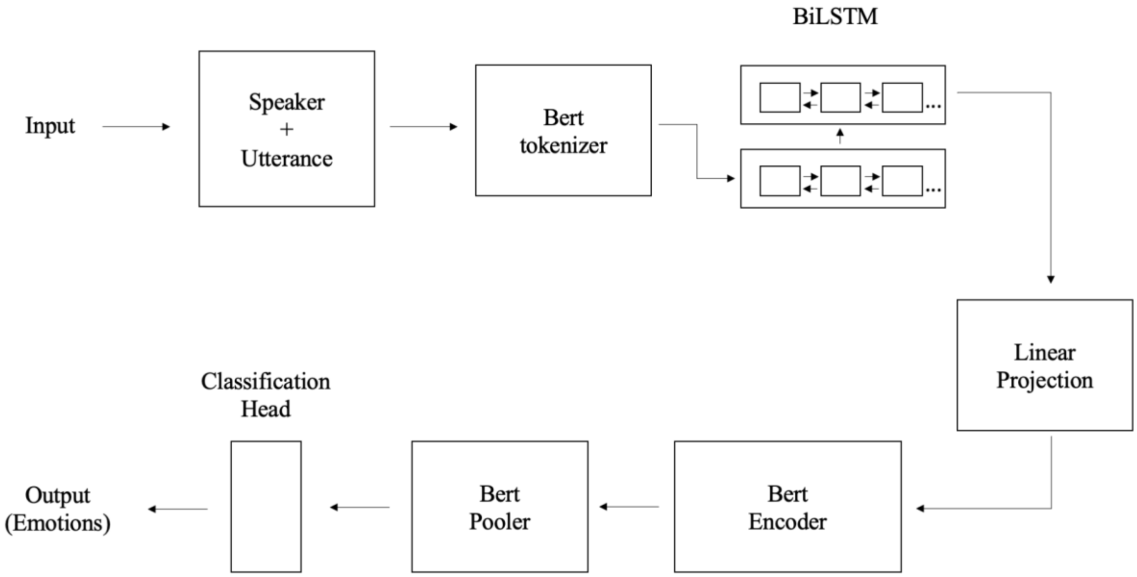
Emotion Recognition in Conversation

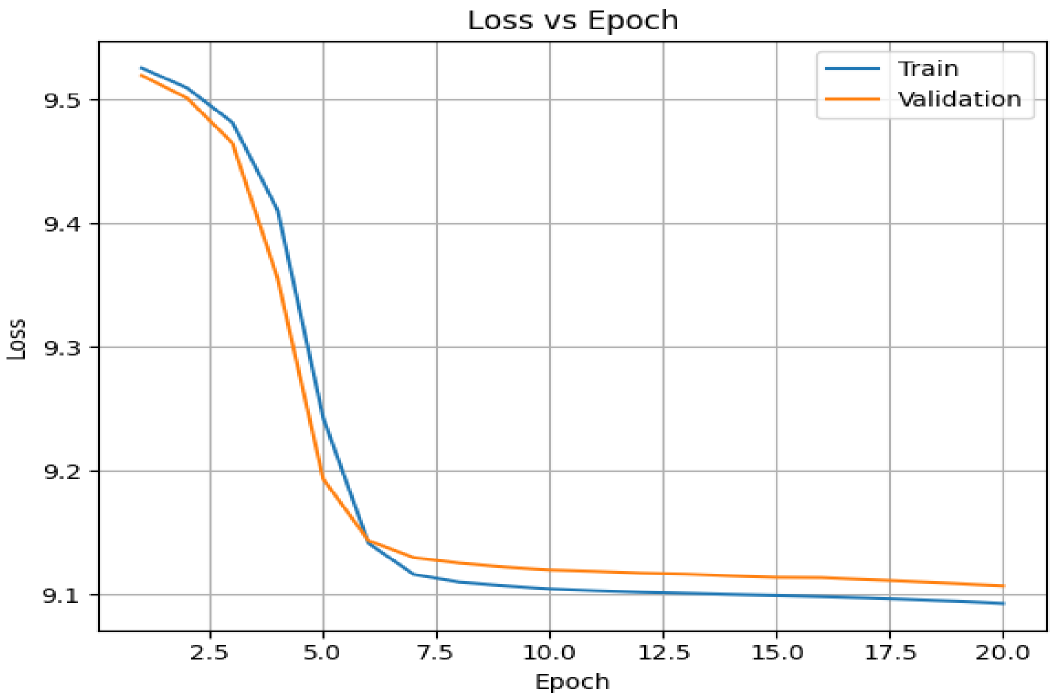
Emotion Recognition in Conversation (ERC) is a specialized field that focuses on automatically identifying and interpreting the emotional states expressed by individuals during conversations. Unlike traditional approaches that analyze emotions in isolated text, ERC aims to understand the nuanced emotional dynamics in conversational exchanges involving multiple speakers. It enables a deeper understanding of emotional processes and interactions, benefiting applications such as conversational agents and affective computing systems.

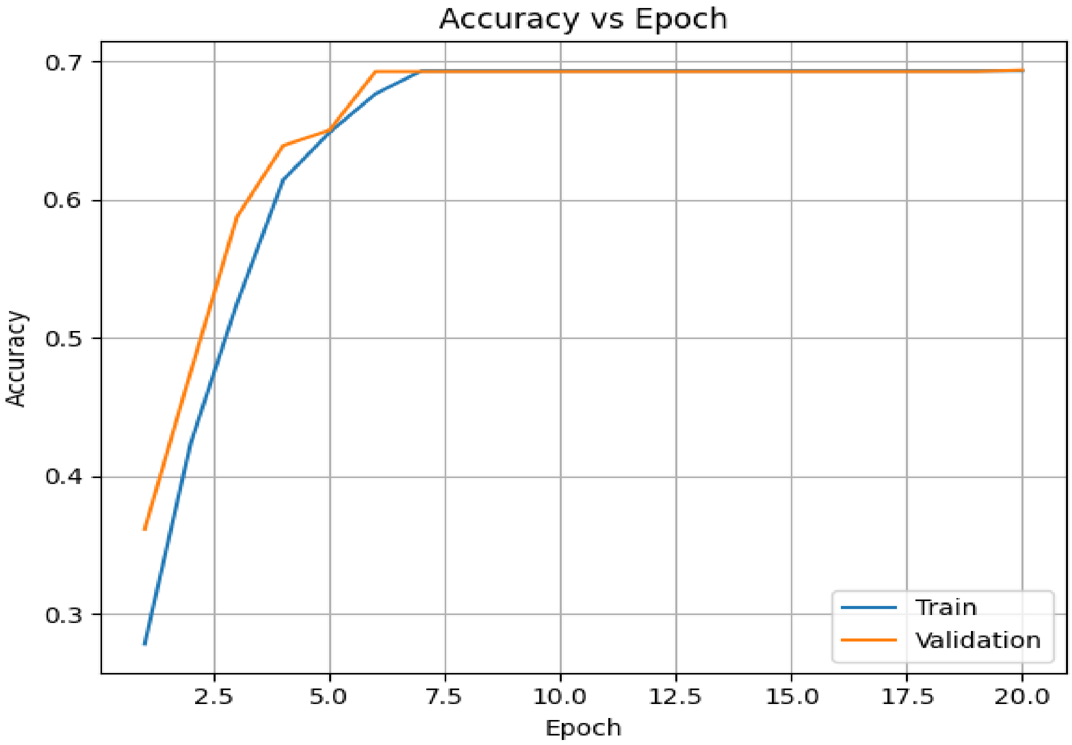
We developed 2 models Model1 and Model2 of which later one performed better. Architecture of both models along with its training and validation accuracy is mentioned below.

Model 1

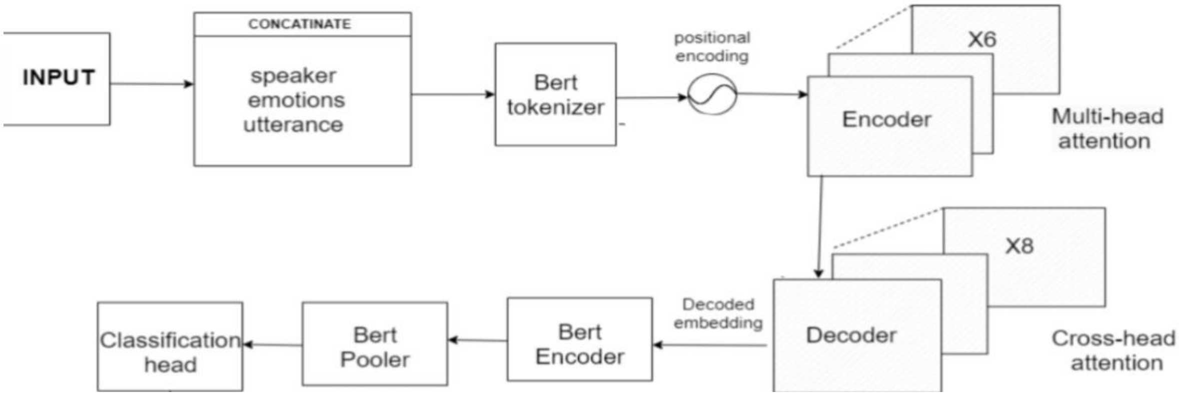


The process begins with transforming the input into numerical values using the BERT tokenizer. These numerical representations are then fed into a Bidirectional Long Short-Term Memory (BiLSTM) model to capture sequential relationships within the data. Next, a linear projection is applied for dimension correction. Following this, the pre-trained BERT encoder and Pooler layers are utilized to further encode the contextual information. Finally, a classification head is employed for prediction

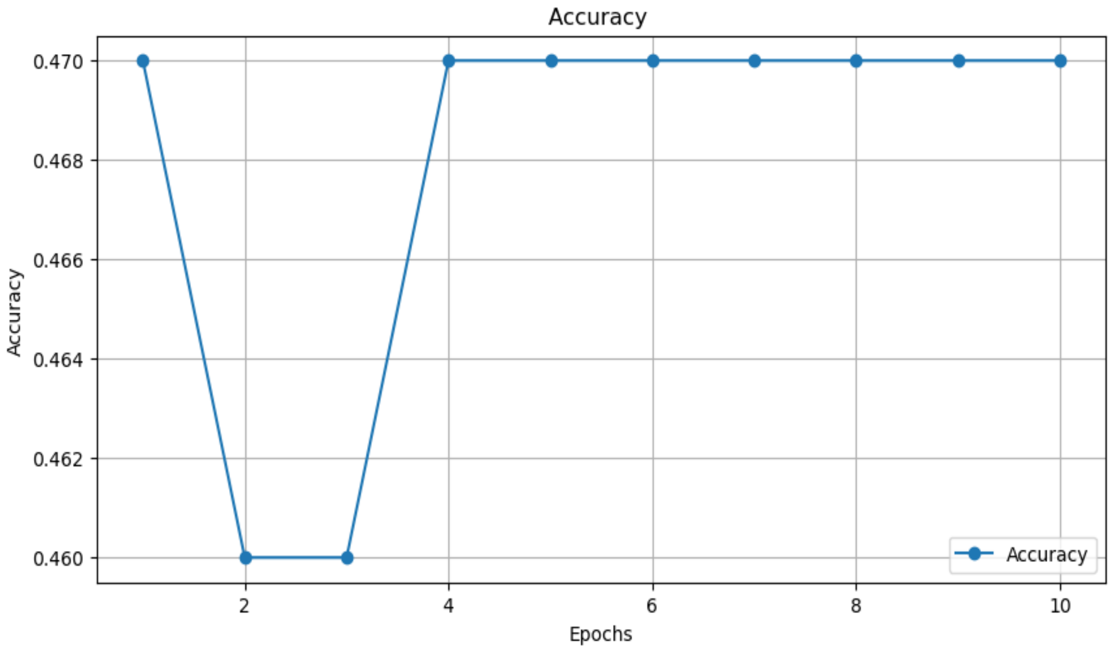

Loss vs Epoch
Train
Validation
Loss
9.5
9.4
9.3
9.2
9.1
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Epoch

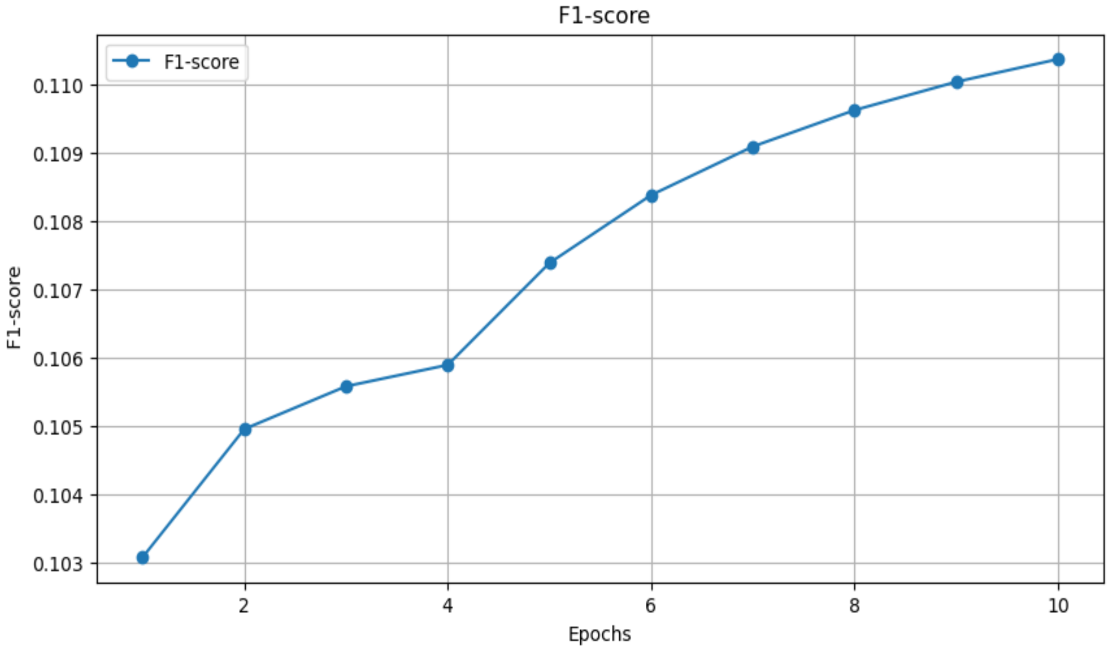

Accuracy vs Epoch
Accuracy
0.7
0.6
0.5
0.4
0.3
Train
Validation
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Epoch

Model 2



The input consists of concatenated speaker, emotion, and utterance information, which undergoes tokenization using the BERT tokenizer to convert it into numerical values while preserving the positional encoding of each word and sentence. These numerical representations are then fed into a transformer encoder to capture relationships through self-attention mechanisms, forming a memory unit. The last sentence's embedding from this memory unit is transferred to the decoder. Here, cross-attention is applied between the last utterance and the memory unit to incorporate contextual information. Finally, leveraging a pre-trained BERT encoder and pooler, along with a classification head, facilitates accurate predictions.



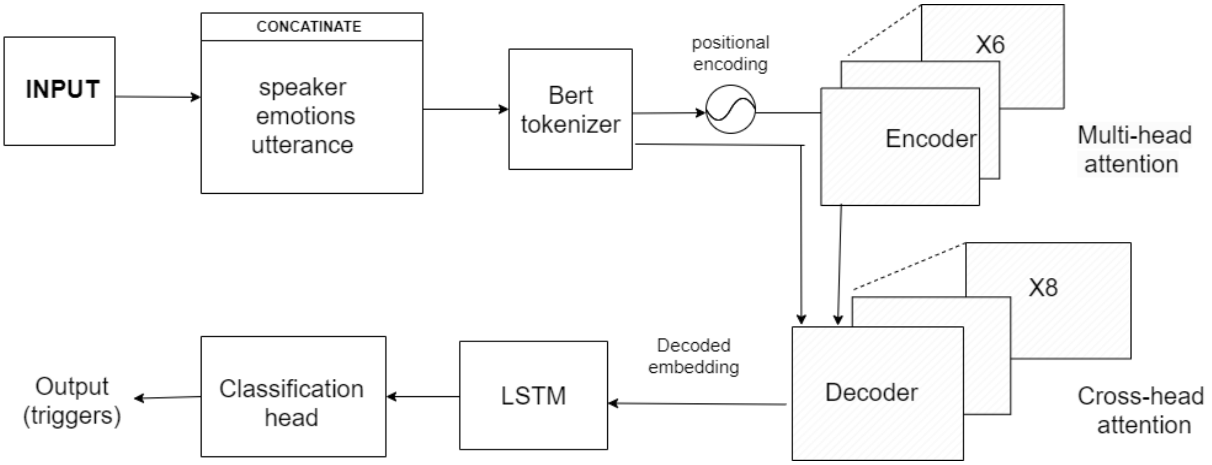


Comparison:

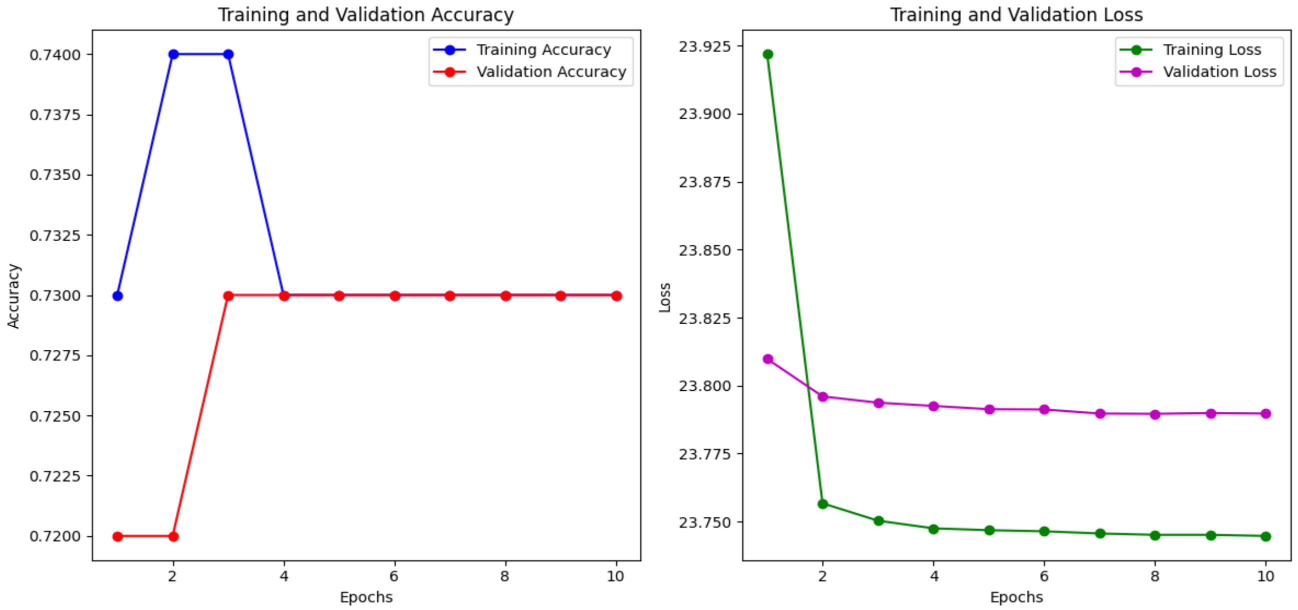
Bi-LSTM models have demonstrated superior capability in understanding and capturing sequential patterns in data, due to their temporal nature. In contrast, Model 1, being less complex, is more adept at handling smaller datasets, where it can generalize effectively without requiring vast amounts of data. On the flip side, the more complex model, characterized by a larger parameter set, demands a substantial dataset to reach its full potential and avoid overfitting.

Emotion Flip Reasoning

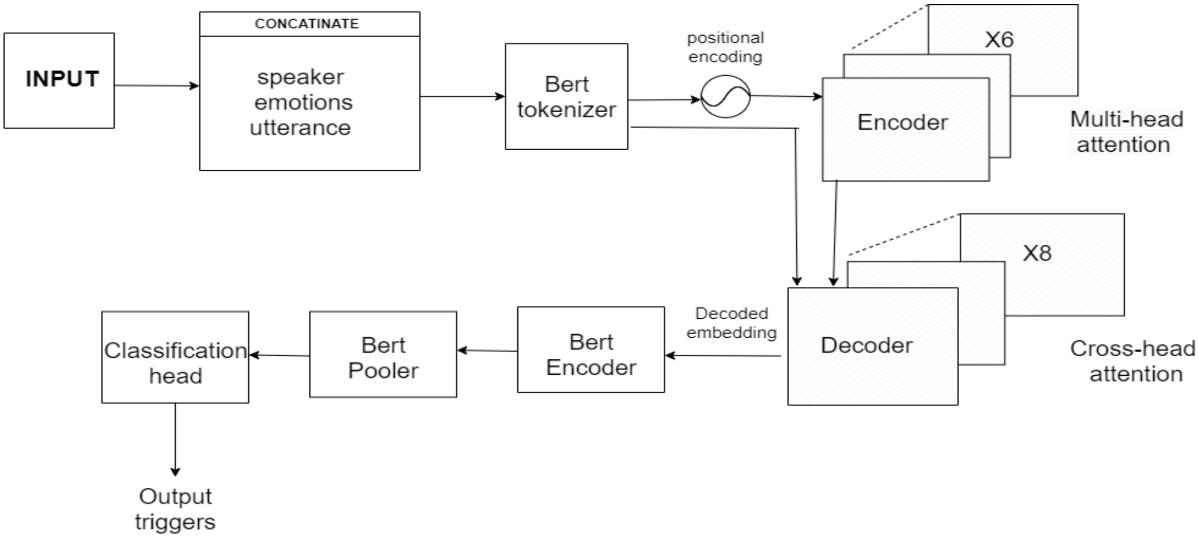
Model 3



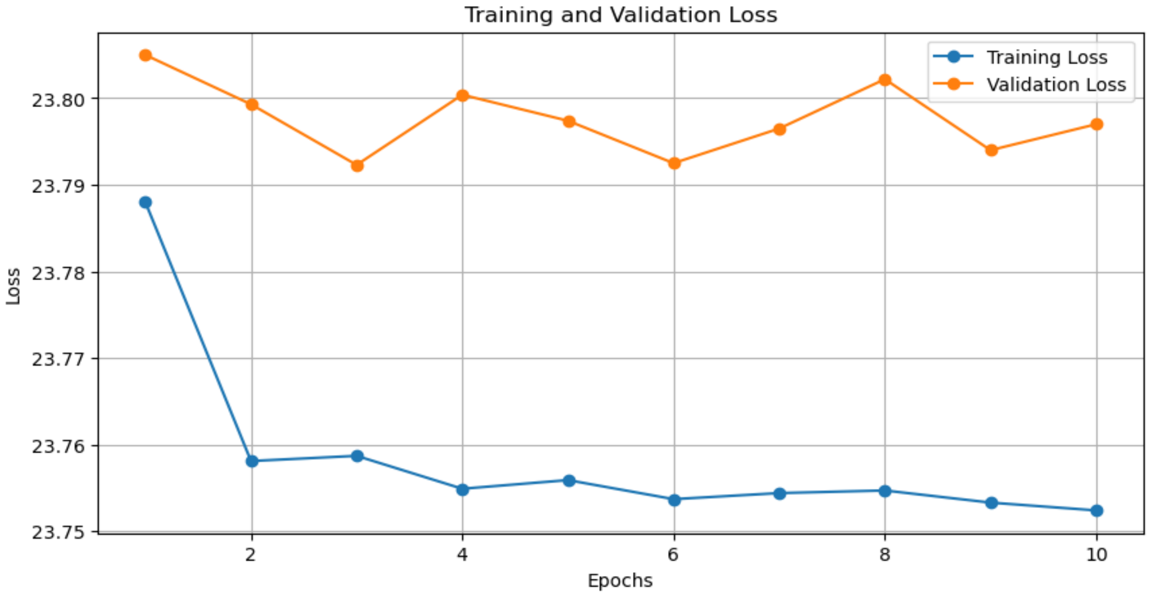
The input, comprising concatenated speaker, emotion, and utterance data, is tokenized using BERT to convert it into numerical values while maintaining positional encoding. These values are then processed by a transformer encoder, enabling the capture of relationships via self-attention, creating a memory unit. The decoder receives the last sentence's embedding from this memory unit, facilitating cross-attention between the last utterance and memory to integrate contextual cues. Additionally, an LSTM is utilized for sequence identification, followed by a classification head for prediction.

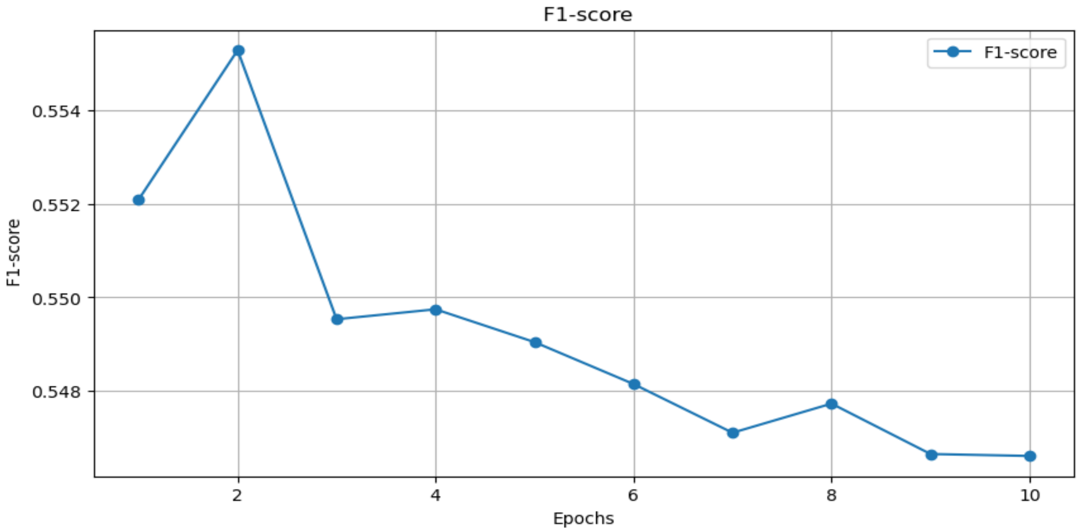


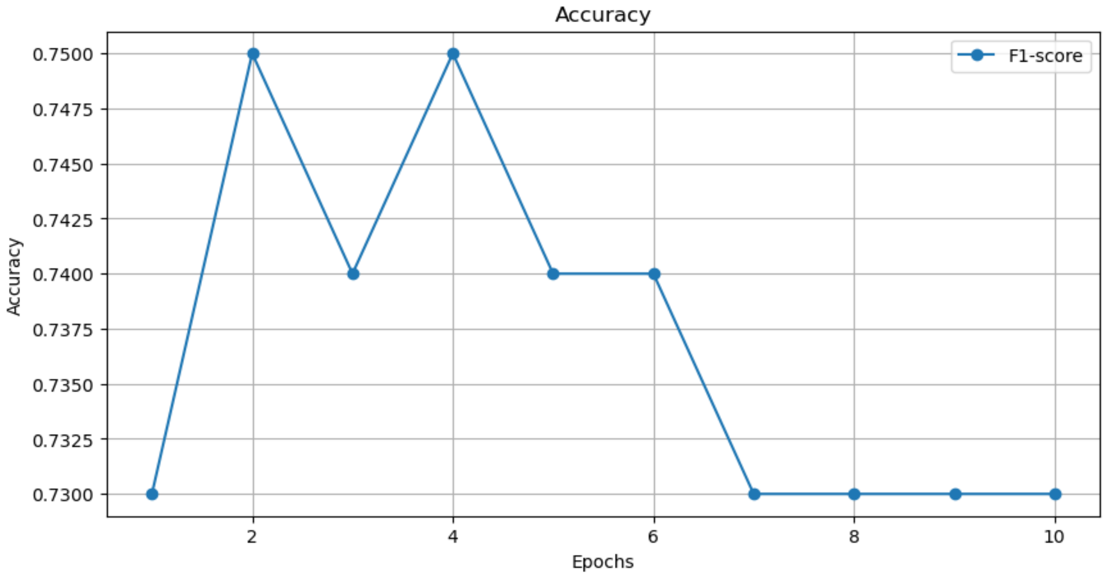
Model 4



The input consists of concatenated speaker, emotion, and utterance information, which undergoes tokenization using the BERT tokenizer to convert it into numerical values while preserving the positional encoding of each word and sentence. These numerical representations are then fed into a transformer encoder to capture relationships through self-attention mechanisms, forming a memory unit. The last sentence's embedding from this memory unit is transferred to the decoder. Here, cross-attention is applied between the last utterance and the memory unit to incorporate contextual information. Finally, leveraging a pre-trained BERT encoder and pooler, along with a classification head, facilitates accurate predictions.







Comparison:

Both models are working almost identical because both are transformer based encoder-attention-decoder architecture. Before classification layer, both the model have different implementation. Former one has LSTM which captures temporal features better than the later one which has BERT encoder and pooler.

Model	Loss (train)	Loss (val)	F1-macro
M1 (Task 1)	0.1710	9.1068	0.1703
M2	8.9142	8.9288	0.1104
M3 (Task 2)	23.7449	23.789	0.5474
M4	23.7524	23.797	0.5466

Contributions (equal):

Model 1: Mohammad Seraj
Model 2: Surabhi Singh
Model 3: Vindhya Regonda
Model 4: Manvendra Nema